8-25-98
B.Hilliard

# IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Re: Patent Application of
F. Herz et al

Docket No.: 6099/002

Serial No.: 08/551,198

Examiner:Huynh-Ba

Filing Date: October 31, 1995

Group Art Unit: 2415

For: SYSTEM FOR CUSTOMIZED ELECTRONIC IDENTIFICATION OF DESIRABLE OBJECTS

OIPE
AUG 17 1998
PATENT & TRADEMARK OFFICE
JC25

BEST AVAILABLE COPY

**CERTIFICATE OF MAILING UNDER 37 CFR 1.10**

I hereby certify that this paper (along with any paper referred to as being attached or enclosed) is being deposited with the United States Postal Service on the date shown below with sufficient postage as Express Mail, Express Mail Label No. EL055940569US in an envelope addressed to Box AF, Assistant Commissioner for Patents, Washington, D.C. 20231

17 AUGUST 1998
Date

James M. Graziano

RECEIVED
98 AUG 20 PM 1:47
GROUP 2700

## RULE 131 AFFIDAVIT

The Assistant Commissioner
for Patents
Washington, D.C. 20231

Dear Sir:

STATE OF )
) ss.
COUNTY OF )

We, Frederick S. M. Herz, a resident of Davis, State of West Virginia; Jason M. Eisner a resident of Philadelphia, State of Pennsylvania; Steven L. Salzberg a resident of Baltimore, State of Maryland; Jonathan M. Smith a resident of Princeton, State of New Jersey; being duly sworn, depose and say:

That we are co-applicants named in the above-titled patent application;

That prior to March 29, 1995, we jointly conceived and subsequently jointly constructively reduced to practice the invention disclosed in the above-identified patent application;

That there are no documented records providing the exact date of conception of the above-identified patent application;

That the invention described in the above-identified patent application was subsequently constructively reduced to practice, as shown in the attached documents Exhibit A and Exhibit B, which documents were produced prior to March

29, 1995;

That the attached Exhibits A and B are reproductions of the original records (with dates blacked out) referred to in this Affidavit;

That the attached Exhibit A comprises a facsimile copy of a draft disclosure document that we created prior to March 29, 1995 and transmitted to James M. Graziano, attorney of record in the above-identified patent application, via facsimile prior to March 29, 1995;

That the attached Exhibit B comprises a draft of the above-identified patent application, produced by the above-mentioned James M. Graziano and transmitted to us prior to March 29, 1995;

That both Exhibit A and Exhibit B describe the basic concept of the invention claimed in the above-identified patent application, that is the concept of providing a user with access to electronically stored target objects via the automatically generated target profiles for the target objects and user target profile interest summaries;

That this novel structure is disclosed in both the attached Exhibit A and Exhibit B, as well as specifically claimed in the above-identified patent application, as in for example claim 1 as it presently stands in prosecution:

> 1. A method for providing a user with access to selected ones of a plurality of target objects that are accessible via an electronic storage media, where said users are connected via user terminals and bidirectional data communication connections to a target server system which includes said electronic storage media, said method comprising the steps of:
>
> automatically generating target profiles for target objects that are stored in said electronic storage media, each of said target profiles being generated from the contents of an associated one of said target objects and their associated sets of target object characteristics;
>
> automatically generating at least one user target profile interest summary for a user at a user terminal, each said user target profile interest summary being generated from target profiles associated with ones of said target objects accessed by said user; and
>
> enabling access to said plurality of target objects stored on said electronic storage media by users via said target profiles and said at least one user target profile interest summary.

That the originals of Exhibit A and Exhibit B are in the possession of James M. Graziano, attorney of record in the above-identified patent application; and

That all of these acts and record were made during the regular course of business in the United States of America prior to March 29, 1995.

We hereby declare that all statements made herein of our own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like are punishable by fine or imprisonment, or both, under 18 U.S.C. §1001, and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Frederick S. M. Herz

Subscribed and sworn to before me this 29 day of June, 1998.

My commission expires: Sept 12, 2000

OFFICIAL SEAL
NOTARY PUBLIC
STATE OF WEST VIRGINIA
REBECCA WHITE
RT. 2, BOX 102
WYMER, WV 26554
My Commission Expires September 12, 2000

(SEAL)

Rebecca White
Notary Public

Jason M. Eisner

Subscribed and sworn to before me this _____ day of _____, 1998.

My commission expires:

Notary Public

(SEAL)

That the originals of Exhibit A and Exhibit B are in the possession of James M. Graziano, attorney of record in the above-identified patent application; and

That all of these acts and record were made during the regular course of business in the United States of America prior to March 29, 1995.

We hereby declare that all statements made herein of our own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like are punishable by fine or imprisonment, or both, under 18 U.S.C. §1001, and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

_____

Frederick S. M. Herz

Subscribed and sworn to before me this _____ day of _____, 1998.

My commission expires:

_____ Notary Public

(SEAL)

Jason M. Eisner

Jason M. Eisner

Subscribed and sworn to before me this 15th day of July, 1998.

My commission expires: 8/13/2001

NOTARIAL SEAL
KIMBERLY A. CONTE, Notary Public
Lower Merion Twp., Montgomery County
My Commission Expires August 13, 2001

Kimberly A. Conte

Notary Public

(SEAL)

Steven L. Salzberg

Subscribed and sworn to before me this 29 day of June, 1998.

My commission expires: 3/1/2001

Notary Public

(SEAL)

Jonathon M. Smith

Subscribed and sworn to before me this 13th day of August, 1998.

My commission expires: 10,2,2001.

Notary Public

ILONA M CASTRO NOTARY PUBLIC STATE OF NEW JERSEY (SEAL)

ILONA M. CASTRO
Notary Public, State of New Je
ID# 2189854
Commission Expires October 2, 2001

Notes for a patent on a Draft
**Method for Customized Information Retrieval Using Personal Profiles**

work by: Fred Herz, Mitch Marcus, Jonathan Smith, Eric Brill, Jason Eisner, Lyle Ungar

## Field of the Invention

The present invention relates to a method for automatically determining which news articles a user is most likely to wish to read from an on-line news source such as the AP news wire or Reuters, and allowing the user to select from those articles. More particularly, the method involves constructing a "profile" for each article based on the frequencies with which each word appears in each article relative to its overall frequency of use in all articles. User profiles are then constructed by observing which articles the users read. Because people have multiple interests, multiple profiles are kept for each user, corresponding to multiple topics of interest. Each user is presented with those articles whose profiles most closely match his or her profiles. User profiles are automatically updated on a continuing basis to reflect each viewer's changing interests. Alternatively, documents are grouped into clusters and menus are automatically generated to allow users to navigate the clusters and locate documents of interest. Documents may be news, electronic mail, or product descriptions, and profiles may include data from structured databases as well as free text. Similar methods can be used to allow users to find people with similar profiles and hence interests for the purpose of commerce or pleasurable discussion.

## Description of the prior art

Researchers in the field of information retrieval have devoted considerable effort to finding efficient and accurate methods of allowing users to select documents of interest from a large set of documents. The most widely used methods are based on keyword matching: the user specifies a set of words which s/he thinks will be in a document and the computer retrieves all documents with contain those words. Such methods are fast, but are notoriously unreliable, as users may not think of the right keywords. Use of logical combinations of words and of wild cards (e.g. [gorilla OR chimp*] and [adolescent* OR teen*) helps reduce, but does not solve this problem.

Starting in the 1960's, an alternate approach was proposed: users were presented with a document and asked if it was what they wanted, or how close it was. Each document was described by a profile: a list of the words in the document or, in more advanced systems, a list of word frequencies in the document. One document is said to be similar to another if the distance between their profiles is small. Similarity of document profiles can be used in document retrieval. A user searching for a information about a certain subject can write a description of what they are looking for. The computer then retrieves documents with profiles similar to that of the request. These requests can then be refined using "relevance feedback", where the user rates the documents retrieved as to how close they are to what is being sought. The computer then uses this information to refine the target profile, and the process is repeated until the user either finds enough documents or tires of the search.

Traditional information retrieval uses relevance feedback for retrieving one or more documents in a single area. (It presents a sequence of documents and use distance from desired or undesired documents to determine what to present next.) In many applications such as the reading of newspapers, users are interested in multiple topic areas, and maintain their interests over time. The method proposed in this patent differs from standard

information retrieval methods in that it keeps multiple profiles of each user, which are maintained and refined over time. The use of such profiles is also extended to include automatic generation of menus, and to the matching of people with similar profiles.

*DISCUSS OTHER PATENTS, PAPERS*

This patent speaks entirely of counting words in a document, but the methods proposed can be trivially extended to counting n-grams of letter. (N-grams are just sequences of N contiguous letters in a document. For example, this sentence contains a sequence of 5-grams which starts "For e", "or ex", "r exa", " exam", "examp" ...) This patent could build upon such methods as, for example in patent # (Marc D.) which describes how such n-gram methods can be used for information retrieval.

A number of researchers have looked at methods for selecting documents of most interest to users. Patti Maes and coworkers at MIT produced the Ringo system for recommending musical selections. Their system requires active feedback from the users (users must specify how much they like or dislike each musical selection, as opposed to this system which determines user interests by monitoring their behavior). Also, Ringo keeps a complete list of users' ratings of music selections and makes recommendations by finding which selections were liked by multiple people. Unlike the system described below, Ringo does not take advantage of any available descriptions of the music, either structured descriptions in a data base or free text such as reviews.

*DISCUSS OTHER Clustering work*

A couple of other research groups have looked at the automatic generation and labeling of clusters of documents for the purpose of browsing through the documents. ? and Marc Dumashek propose a method where documents can be displayed on a two dimensional plot with distances between the documents indicating how dissimilar they are. They general cluster lables by ___. A group at Xerox Parc has developed a method they call "scatter/gather" in which ...

The method proposed below differs in that it uses a hierarchical clustering method to generate a menu tree and then further modifies this tree based on data collected as users access data through the menus generated from the tree.

**Overview of patent**

This patent proposes a fundamental methodology for matching people and objects by calculating, using and automatically updating profiles describing the users' interests and the objects' characteristics. The "objects" may be text documents, purchasable items, or even other people. Examples include a person looking for a newspaper story of potential interest, a movie to watch, an item to buy, or another person to correspond with. In all cases, the method is based on determining the similarity between a profile for the object and a profile for the user. Object profiles may include some or all of the following: (1) structured text (as in a newspaper story, a movie review, a product description or an advertisement), (2) structured data (e.g. the author of the newspaper article, the actor and directors and rating given to a movie, the price and manufacturer or a product, or the name and phone number of the person placing an advertisement) and (3) a list of the persons who have accessed (e.g. read or purchased) the object. The structured data may, in particular, contain information about the quality of the object, such as measures of its popularity (how often it is accessed) or of user satisfaction (number of complaints received).

The ability to measure the similarity of profiles describing objects and similar profiles describing a user's interests can be applied in two basic ways: filtering and browsing.



2

Filtering is useful when large numbers of documents are being received by a user, who only has time to read a small fraction of them. For example, one might potentially receive all items on the AP news wire service, all items posted to a number of news groups, or all advertisements in a set of news papers - but few people have the time or inclination to read so many documents. A filtering system (described below) selects a subset of the documents which the user is more likely to wish to read. The retrieval method described below improves its filtering accuracy over time by noting which documents the user reads.

Browsing provides an alternate method of selecting a small subset of a large number of documents. (We use the word "documents" below, but please keep in mind that the documents could be descriptions of items being sold or of different people and their interests.) Documents are organized so that users can navigate among the documents by moving from broad groups of documents to smaller, more specific, groups, to individual documents, or from documents or document groups to closely related documents or groups. The methods presented below allow documents to be grouped into clusters and the clusters to be grouped into larger and larger clusters. These hierarchies of clusters then form the basis for menuing and navigational systems to allow rapid search large numbers documents.

The following sections discuss a number variations on the theme of developing and using profiles for document retrieval:

the basic implementation of an on-line news clipping service,
a variation to do filtering of electronic mail,
an extension for retrieval of objects such as purchasable items which may have more complex descriptions,
a method of automatically building and altering menuing systems for browsing, and
a method of constructing "virtual communities" of people with common interests

The final section between the claims describes a computer architecture which permits these filtering and retrieval methods to operate efficiently when the information is distributed across dozens, hundreds, or millions of different machines.

## Preferred embodiment - news clipping service

This section describes the technology underlying all the methods revealed in this patent, by showing how they are used in an automatic news clipping service which learns to select (filter) news articles to match a user's interests, based solely on which articles the user chooses to read.

Each document is assigned a profile based on the relative frequencies of occurrence of the words in the document. Users are also assigned profiles by a method described below. As new documents are received, their profiles are compared to the user's profiles, and documents which are closest (most similar) to the user's profile are presented to the user for possible reading. The computer program providing the articles to the user monitors how much the user reads (the number of screens of data and the number of minutes spent reading), and adjusts the user's profiles to more closely match what was read.

See figure 1.

1. retrieve new documents from document source (e.g. news from AP news wire)

2. calculate document profiles

3. compare document profiles to user profiles; select closest documents

4. present list of documents to user

5. monitor which documents are read

6. update user profiles

**Fig. 1 Method for selecting documents of most interest to computer users**

Each of the steps outlined in figure 1 is described in detail below. The details of the method require selecting means of calculating profiles, of measuring similarity of documents and profiles, and of updating profiles based on what the user read.

1. Retrieve new documents from document source.

Documents are available on-line from a wide variety of sources. In the preferred embodiment, one would use the current days news as supplied e.g. by the AP or Reuters news wire. These methods are equally useful for selecting which articles to read from news groups and electronic bulletin boards, and can be used as part of a system for screening and organizing electronic mail ('email').

2. Calculate document profiles.

A profile is computed for each document which indicates the relative frequencies of word occurrences in this document relative to other comparable documents. Many measures can be used, but the preferred profile is calculated using the TF/IDF measure: the term (word) frequency (TF) times the inverse document frequency (IDF), where the word frequency is the number of occurrences of a word in a document divided by the total number of words in the document and the inverse document frequency is taken to be one over the logarithm of the fraction of documents in which the word occurs. (Other measures can be used, but generally do not work as well.) Words which occur frequently ("a", "and", "the" ...) influence calculated document similarities without indicating document topics, and so are typically removed from feature vectors. Those skilled in the art will be familiar with the large literature on this subject; see, for example Salton and McGill (1983).

Other methods of calculating relative word frequencies can equally well be used, such as latent semantic indexing or probabilistic models (refs: Rij..., etc.) For many applications it is useful to augment the set of words actually found in the document with a set of synonyms or other words which tend to co-occur with the each word in the document. This provides profiles which match a broader class of documents and reduces the chance of missing desired documents.

Synonym dictionaries which group together similar words can be used to improve the performance of the profiling and menuing systems. Words which are similar can be treated

as if they were the same word, or documents containing a word can be augmented with synonyms of the word or with other words that tend to occur with the word. Thus an article talking about "staples" could also be matched to articles about "staplers". The stapler example also illustrates the potential utility of morphological analysis where "staple", "stapled" and "staples" would be treated as the same word.

*talk about hash coding, efficiency ...*

3. Compare document profiles to user profiles; select closest documents.

A set of profiles is stored for each user. These are initially selected by using the profiles of documents that the user indicates are representative of his interest, by asking the user for key words, or by using a standard set of profiles for the people in a given demographic mix. All new documents are compared against each of the user's profiles, and those documents closest to each profile are selected. We use the cosign measure of similarity between documents and queries:

$$\text{cosine}(\text{Profile}_i, \text{Profile}_j) = \frac{\Sigma\,(\text{term}_{ik}\,\text{term}_{jk})}{\text{sqrt}(\Sigma\,\text{term}_{ik}^2\,\Sigma\,\text{term}_{jk}^2)}$$

where all sums are over k - sums over all the terms (TF/IDFs for each word), $\text{term}_{ik}$ is the TF/IDF for the kth word in profile i, and $\text{term}_{jk}$ is the TF/IDF for the kth word in profile j. (Since most words do not appear in most documents, most terms in most document profiles are zero, the documents are almost orthogonal to each other, and obvious measures of the similarity of between documents such as Euclidean distance do not work well.) Other similarity metrics such as the dice measure can also be used.

4. Present list of documents to user.

Titles of the selected documents are presented to the user, who can then select any document for viewing. (If no titles are available, then the first sentences of each document can be used.) The list of documents is ordered according to the similarity of the documents to the user's profiles. Users can still access all documents; those farthest from the users interests are simply lower on the list.

5. Monitor which documents are read.

The system monitors which documents the user reads, keeping track of how many pages of text are read, how much time is spent viewing the document, and whether all pages of the document were viewed. This information can be combined to measure the user's interest in the document. Although the exact details depend on the length and nature of the documents being searched, a typical formula might be

measure of document attractiveness =

- 0.2 if the second page is accessed +
- 0.2 if all pages are accessed +
- 0.2 if more than 30 seconds was spend on the document +
- 0.2 if more than one minute was spend on the document +
- 0.2 if the minutes spent in the document are greater than half the number of pages

6. Update user profiles.

Updating of user profiles is done using the method described in patent (our Video Patent). The user profile closest the document read is shifted slightly towards the document profile.

*INSERT 3 PARAGRAPH DESCRIPTION OF PATENT*

< add alternate embodiments and Other considerations.>

**Preferred embodiment - Specialization for electronic mail (email)**

*FROM ERIC BRILL*

**Extension: Matching users for purchasing and virtual communities**

The user profiles described above can also form the basis for matching buyers and sellers, people bartering goods, and people with common interests. These matches may be either pairs of people or corporate entities, as in a buyer and a seller, or they may be larger groups, as in virtual communities of people wishing to discuss subjects of common interest.

It is common for merchandisers to provide catalogues with descriptions of the products which they sell. These are now often available on-line on CD-ROMS or the internet. The profile-based news clipping and information retrieval techniques described above can also be used to help users locate items which they may wish to purchase. Profiles of the product descriptions and of product reviews can be made, treating them as standard documents augmented with structured database information. Users repeatedly shopping for given items can have profiles created and modified to reflect the item descriptions which they find most interesting. For single purchases, menus generated by clustering items can be used. This is particularly important when a user may be shopping from items sold by a large number of vendors, where vendors' descriptions and product categories are variable.

Such profiles can also be used to match any two people based either on profiles generated from descriptions they provide ("want ads" to buy or to sell) or based on profiles of what they have purchased or sold in the past. Such matching is particularly critical in "thin" markets such as most barters, and in certain contract employment situations, where those involved in the barter or contracting are looking for specific items or abilities. In a market where many different people are looking for unusual products or services, the ability to automatically search through millions of product and service descriptions to find matches is particularly valuable. As before, the filtering (news clipping-style profile updating) is useful when one is repeatedly seeking similar people (e.g. an ongoing need for database consultants with specific experience), while the browsing (cluster-based menu) system is more useful for unusual searches (e.g. one is looking for a consultant for a specific data base job, based on a description of that job).

If the descriptions of the purchasable items are free text, then the methods described for the news clipping service can be used directly. However, purchasables often have structured descriptions, including attributes such as price, availability of colors or sizes, delivery times and costs, popularity (volume sold), ratings by users or independent organizations (e.g. Consumer Reports), etc. Such items can be included along with relative word frequencies (e.g. TF/IDF measures) in the document profiles, as long as care is taken to accurately compute the relative weightings to be given to the different attributes. <see Home video patent again>

**Preferred embodiment • virtual community**

Computer users frequently join other users for discussions on computer bulletin boards or news groups. In current practice, each bulletin board has a specified topic and users then look through a long list of topics (typically hundreds) to find topics of interest. The users then must select for themselves which of thousands of messages they find interesting from among those posted to the selected bulletin boards and, if desired, post additional discussion on the topics. The profiling system described above can also be used to allow users to find other people with similar interests (i.e. to generate virtual communities of people with similar interests). Profiles of each person can be formed by the news-clipping method described above: a person's set of profiles is basically the centroids of the clusters of the profiles of the bulletin board articles read by the user. Users can then be grouped together based on the similarity of one or more of their profiles. Such groups of people have been reading and writing about similar topics and in similar styles and so will presumably share interests. New bulletin boards can be formed as sufficient numbers of users with different interests accumulate.

The existence of thousands of Internet bulletin boards (also called news groups) and countless more private bulletin board services (BBS's) demonstrates the very strong interest among members of the electronic community in forums for the discussion of ideas about almost any subject imaginable. Currently, bulletin board creation proceeds in a haphazard form, usually instigated by a single individual who decides that a topic is worthy of discussion. There are protocols on the Internet for voting to determine whether a news group should be created, but there is a large hierarchy of news groups (all beginning with the prefix "alt.") that do not follow this protocol.

One can construct a browser for bulletin boards where each bulletin board is characterized by one or more profiles, so that potential new bulletin board users can locate bulletin boards by presenting the system with one or more messages typical of what they are looking for. Profiles would be generated from the messages using the methods described above, and then the user pointed to those bulletin boards most closely corresponding to the messages. Note that bulletin boards, like all objects with profiles, could also be clustered and put into an automatically generated menu.

The above method will be useful if people with the right set of interests have already formed a bulletin board or "chat group". However, because people have varied and varying complex interests, it is desirable to automatically create groups of people ("virtual communities" with common interests. The Virtual Community System (VCS) described below is a network-based agent that seeks out users of a network with common interests, dynamically creates bulletin boards or electronic mailing lists for those users, and introduces them to each other electronically via e-mail.

The functions of VCS are as described below. These are general functions that could be implemented on any network ranging from an office network in a small company to the World Wide Web or the Internet. The four main steps in the procedure are as described in Figure 3:

1. Scanning postings to bulletin boards and clustering people who post similar messages.

2. Creating new bulletin boards

3. Announcing the bulletin boards

4. Enrolling people in the bulletin boards

**Fig. 3 Method for creating virtual communities**

Each of these steps is carried out as follows:

**Scanning.** VCS finds people with interests in common who form a virtual community. Using the text-searching and clustering technology described above, the VCS will constantly scan all the news groups and electronic mailing lists on a given network. This could be the Internet, or a set of bulletin boards maintained by America On-Line, Prodigy, or Compuserve, or a smaller set of bulletin boards that might be local to a single organization, for example a large company, a law firm, or a university.

Clustering the messages sent to bulletin boards based on the similarity of their profiles automatically finds threads of discussion that show common interests among the users. Naturally, discussion on a single bulletin board will tend to show common interests; however, this method uses all the texts from every available bulletin board and electronic mailing list. Whenever it finds a cluster of sufficient size (generally 10-20 different messages) in different bulletin boards or mailing lists, it will form a new virtual community.

**Bboard creation.** VCS creates a bulletin board or an electronic mailing list, whichever is appropriate, representing the newly-formed community. If the newly-found cluster only contains a small number of members, for example 2-10 people, it is probably most appropriate to create a small mailing list and subscribe these people to it. VCS will initiate the mailing list by sending an e-mail message containing the text it found and a suggested name (see below for techniques for generating cluster labels) for the new mailing list. After this initial message, users may choose to respond and continue the discussion, or they may let it expire if they prefer. All such mailing lists will be time-stamped, and if they are not used for a user-determined length of time, VCS will delete them.

If VCS finds a larger number of people engaged in a discussion in different forums, it will create a new bulletin board instead. (The number required to create a bboard can be set by the person installing VCS on the network.) It will create a name for this bulletin board and post on it the messages it has collected; i.e., all the messages in a cluster. For a short period of time (probably a few weeks), it will continue collecting messages automatically using its clustering routines, and these messages will be posted to the bulletin board. After the time limit has expired, new readers of the bulletin board will have to maintain it by posting items voluntarily. Alternatively, the system manager can allow VCS to continue posting items automatically for an indefinite period of time, if this is deemed useful.

**Announcing.** VCS informs all the members of the new virtual community of the new bulletin board service (or mailing list), and gives each of them a brief summary of the basis of the community. Because bulletin boards are read voluntarily, VCS will send an e-mail message directly to everyone who posted any text that served as the

basis for a new bulletin board it has created. As stated above, it will do the same for new mailing lists. However, with bulletin boards, it will simply inform users of the existence and name of the bulletin board, and leave it to them to decide whether or not they wish to read it.

**Enrolling.** Even after creation of a new Virtual Community, VCS will continue to scan existing mailing lists and bulletin boards for messages that belong to that community. Any new messages that appear will be cross-posted to the new community, and the person posting the message will be informed by VCS that he/she belongs to a new Virtual Community. The user can then decide whether or not to read the new VCS bulletin board or, if the community is using a mailing list, to subscribe to the mailing list.

With these facilities, VCS will provide automatic creation of virtual communities in any local or wide-area network. The core technology underlying VCS is creating a search and clustering mechanism that can find documents that are "similar" in that the users share interests. This is precisely what was described above. One must be sure that VCS does not bombard users with notices about communities in which they have no real interest. On a very small network a human could be ``in the loop," scanning proposed virtual communities and perhaps even giving them names. But on larger networks the VCS has to run in fully automatic mode, since it is likely to find a large number of virtual communities.

**Preferred embodiment - Menu generation based on clustering and automatic cluster labeling**

A browsing system can be constructed for the retrieval of other documents such as reference articles in an on-line library using the same methods underlying the news clipping service. The main difference is that since all reference articles are potentially of interest (not just the most recent ones, as is the case for news), and since users are often looking for specific topics, the preferred interface includes a menuing system.

In order to allow users to rapidly locate a document from among a large set of documents, we first calculate profiles for each document using the TF/IDF word frequency measures described above. Documents are then grouped into clusters using a hierarchical clustering algorithm such as a k-means clustering algorithm. Clusters are groups of documents which are similar to each other, i.e. close to each other in some metric such as the cosign measure. Hierarchical clusters produce a tree which divides the documents first into two large clusters of roughly similar documents; then each of these clusters is in turn divided into two or more smaller clusters, which in turn each divided into yet smaller clusters until a "cluster" is found consisting of a single document. This division of documents provides an efficient method to retrieve a document of interest: the user first chooses between the highest level (largest) clusters and then selects among the smaller clusters. This process is repeated until the user comes to the lowest level in the tree, which are the documents themselves.

Hierarchical trees allow rapid selection of one document from a large set. In ten menu sections from menus of ten items each, one can reach $10^{10}$=10,000,000,000 (one hundred billion) items.



9

*add description of figure:*

1. calculate document profiles
2. cluster documents into a hierarchical cluster
3. generate labels for each cluster
4. generate menus from cluster structure and labels
5. optionally, monitor which documents are read and adjust menus

**Fig. 2 Method for selecting documents using menus**

## Label generation

A key requirement to successfully use the menu is the ability to automatically label the clusters. Labels can be generated by selecting the document or documents closest to the center of the cluster and then displaying either the document title, or the set of words in the profile of the centroid of the document cluster which have the highest relative frequency (TF/IDF). More informative labels can be generated by (1) removing redundant synonyms from the labels either using a synonym dictionary or a morphological analyzer (e.g. "salt, salted, salty, saltier...." should reduce to "salt") and (2) using terms which have the highest discriminatory ability as labels, rather than those with the highest frequency. *<explain how this is done>*

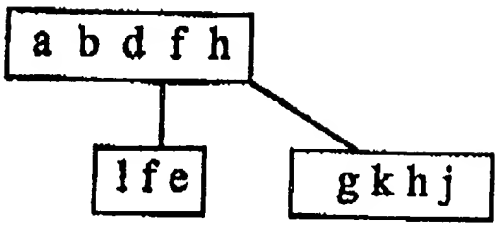
It is also useful to allow users to view the documents as needed in case the above descriptions are insufficient.

## 4 Menu generation

Although most clustering methods generate binary branching trees, for human users, it is preferable to have 4-8 items in a menu. This is easily accomplished by displaying the four grand children or eight grand children of a node in a cluster. (see figure x.)

## 5 Menu updating

As users access documents using the menuing system, certain documents and certain regions of the document profile space will be accessed more frequently than others. These regions correspond to the users interests. If, for example, a user frequently accessed documents close to a, b, and d in figure 3(a) then the menu in 3(d) could be modified to show



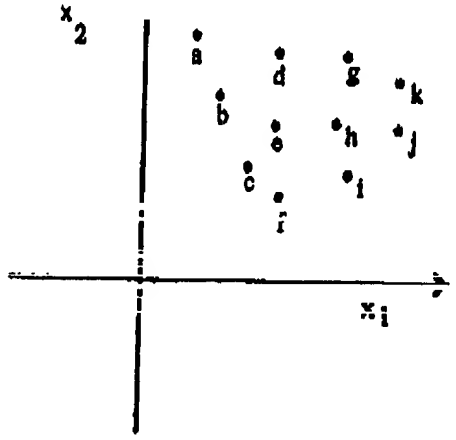

could use decision tree

A more formal algorithm for this is: associate with each node of the tree a probability that the document the user selects will be located under that mode (i.e., is in that cluster). These probabilities can be either all set equal to the number of documents in the cluster divided by the total number of documents, or (if data are available) can be based on total access to each document by all users. (I.e., the probability of access of a cluster is the total number of accesses of the document in the cluster divided by the total number of accesses of all documents.)

Once weightings are known for each cluster at each level of the tree, user menus can be generated to maximize the retrieval efficiency. If menus are chosen so that each menu item has approximately equal probability of accessing the documents under it, then the user will have to go through the minimum number of menus.

---

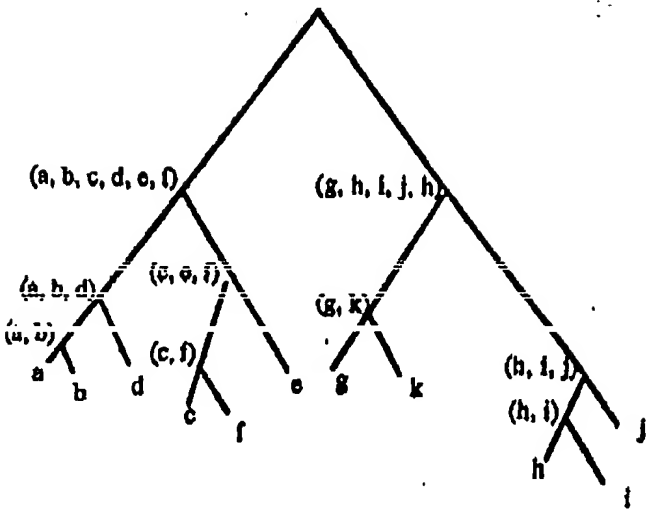
**Figure 3 Menu generation from document clusters**



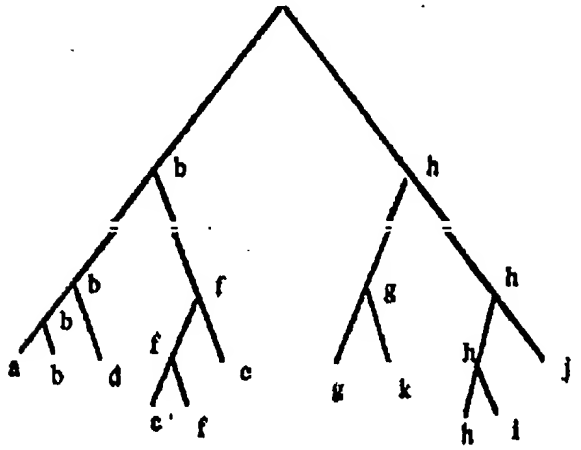

(a) Document profiles. Each letter represents a document. Axes x1 and x2 are two of the hundreds of relative word frequencies. Circles represent clusters.



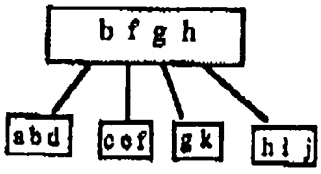
11




(b) Hierarchical cluster tree for documents in (a).




(c) Cluster labels for tree in (b) Document titles, most distinguishing or most frequent words could also be used for cluster labels.




(d) Collapsed menu tree from tree in (c)

**Alternate embodiments of the menuing system**



12

Many variations and improvement on the above menuing system are possible. Different clustering methods can be used, such as fuzzy clustering systems which allow a given document to appear in more than one cluster. When retrieving documents which contain structured information as well as free text, the word frequencies can be supplemented with other terms such as the manually assigned category of the article, the price of the object being described in the document, or expert ratings of the object being described in the document (e.g. number of stars given to a movie in a movie review). Similarity measures (distance metrics) can be adjusted to reflect different stress laid by different users on different terms or items.

The automatic menuing system described above can be used in parallel with manually generated classifications of articles when such are available.

Users can index into clusters either starting at the top of the tree and moving to more specific subclusters or by starting by giving an English language query which is matched against one of the clusters. After an initial cluster is located by finding the cluster center most similar to the query, the user can move to adjacent clusters. It is generally less efficient to start at the largest cluster and repeatedly select smaller subclusters than it is to write a brief description of what one is looking for (e.g. an inexpensive microwave with 1000 watts power, and then to move to nearby clusters if the objects initially recommended are not those desired. <explain further?>

**Preferred embodiment - architecture**

These retrieval methods can also be used on documents stored at multiple locations, such as files at many internet sites. The methods described above still apply, but with minor differences in implementation: software agents profile documents, ideally on each local machine, then the profiles are sent along with pointers to the original documents, to a central machine, where the clustering is done. Documents could be World Wide Web (WWW) pages, general computer files, directories of files, news group postings, e-mail, on line design documents, technical papers on their abstracts, etc.

*FROM JONATHAN SMITH*

**Summary**

A method has been presented for automatically selecting articles of interest to a reader. The method generates profiles of the readers based on the relative frequency of occurrence of words in the articles which they read. It is characterized by passive monitoring (users do not need to explicitly rate the articles), multiple profiles per person (reflecting interest in multiple topics) and use of elements of the profiles which are automatically determined from the data (e.g. the TF/IDF measure based on word frequencies and descriptions of purchasable items).

A method has also been presented for automatically generating menus to allow users to locate and retrieve information on topics on interest. This method clusters documents based on their similarity, as measured by the relative frequency of word occurrences. Clusters are labeled either with document titles or with key words extracted from the document. The method can be applied to large sets of documents (millions to billions) distributed over many machines.

The above methods can be applied to anything for which profiles can be generated, these include news articles, reference or work documents, electronic mail, product or service descriptions, people (based on the documents they read or the descriptions of the products they buy), and electronic bulletin boards (based on the documents posted to them).

**Claims:**

1) A method for automatically building profiles of people reading documents in order to retrieve documents of greater interest to the readers. (E.g., a custom news clipping service)

2) A method for automatically labeling clusters of documents in order to allow users to rapidly locate and retrieve documents on topics of interest to the readers.

2a) A method for screening email based on comparison of user profiles with those of the mail messages.

3) A method for generating menus from hierarchical clusters consisting of the following steps:

4) A method for automatically customizing menu trees to allow users to more rapidly access documents in repeatedly accessed areas of interest.

5) A method for locating products or services based on profiles. Profiles are generated for each product or service based on the word frequencies of words in the product description and reviews and on other descriptive data. Then (a) products or services which the user often seeks information about can be "clipped" for the user as in claim 1 or (b) products and services can be clustered and compiled into a menu as in claim 2-3.

6) A method for matching up different people with common interests for purposes such as buying, selling or bartering goods or services. Profiles are generated for each individual and then individuals are matched based on similarity of their profiles. Again, either a) individuals similar to those who the user often seeks information from can be "clipped" for the user as in claim 1 or (b) individuals can be clustered and compiled into a menu as in claim 2-3.

7) A method for matching of sets of people with common interests ("virtual communities") based on profiles developed from the messages which the people read and send.

8) A scaleable method for retrieving documents distributed over large numbers of computers. <to be completed by JMS>

B

# SYSTEM FOR CUSTOMIZED INFORMATION DELIVERY

## CROSS-REFERENCE TO RELATED APPLICATIONS

This patent application is related to U.S. Patent Application Serial No. 08/??, filed November 28, 1994 and titled "SYSTEM AND METHOD FOR SCHEDULING BROADCAST OF AND ACCESS TO VIDEO PROGRAMS AND OTHER DATA USING CUSTOMER PROFILES", which application is assigned to the same assignee as the present application.

## FIELD OF THE INVENTION

This invention relates to customized information retrieval and delivery in an electronic media environment and, in particular, to a system that automatically constructs both an "article profile" for each article in the electronic media based on the frequency with which each word appears in the article relative to its overall frequency of use in all articles, as well as a "user profile" for each user based on which news articles a user is most likely to wish to read. The system then processes the two profiles to generate a user customized rank ordered listing of articles most likely of interest to the user so that the user can select from these relevant articles automatically selected by the system from the plethora of articles available on the electronic media.

## PROBLEM

It is a problem in the field of electronic media to enable a user to access information of relevance and interest to the user without requiring the user to expend an excessive amount of time and energy. Electronic media, such as on line information sources, provide a vast amount of information to users, typically in the form of "articles", each of which comprises a publication item or document that relates to a specific topic. The difficulty with electronic media is that the amount of information available is overwhelming and the articles that contain this information are not organized on article repository systems that are connected to the electronic media in a manner that simplifies access by a user to only the articles of interest to the user. Presently, a user either fails to access relevant articles because they are not easily identified or expends a significant amount of time and energy to

conduct an exhaustive search of all articles to identify those most likely of interest to the user. Furthermore, even if the user conducts an exhaustive search, present information searching techniques do not necessarily accurately extract only the most relevant articles, but also present articles of marginal relevance due to the functional limitations of the information searching techniques.

Researchers in the field of information retrieval have devoted considerable effort to finding efficient and accurate methods of allowing users to select articles of interest from a large set of articles. The most widely used methods of information retrieval are based on keyword matching: the user specifies a set of keywords which the user thinks will exclusively be found in the desired articles and the information retrieval computer retrieves all articles which contain those keywords. Such methods are fast, but are notoriously unreliable, as users may not think of the right keywords or the keywords may be used in articles in an irrelevant or unexpected context. As a result, the information retrieval computers retrieve many articles which are unwanted by the user. The logical combination of keywords and the use of wild card search parameters help improve the accuracy of keyword searching but do not completely solve the problem of inaccurate search results.

Starting in the 1960's, an alternate approach to information retrieval was proposed: users were presented with an article and asked if it contained the information they wanted, or to quantify how close the information contained in the article was to what they wanted. Each article was described by a profile which comprised either a list of the words in the article or, in more advanced systems, a list of word frequencies in the article. Since a measure of similarity between articles is the distance between their profiles, the measured similarity of article profiles can be used in article retrieval. For example, a user searching for information on a subject can write a short description of the desired information. The information retrieval computer generates an article profile for the request and then retrieves articles with profiles similar to the profile generated for the request. These requests can then be refined using "relevance feedback", where the user rates the articles retrieved as to how close the information contained therein is to what is desired. The information retrieval computer then uses this relevance feedback information to refine the request profile and the process is repeated until the user either finds enough articles or tires of the search.

A number of researchers have looked at methods for selecting articles of most interest to users. An article titled "Social Information filtering: algorithms for automating 'word of mouth'" was published at the CHi-95 Proceedings by Patti Maes et al and describes the Ringo information retrieval system which recommends musical selections. The Ringo system requires active feedback from the users - users must manually specify how much they like or dislike each musical selection. The Ringo system maintains a complete list of users' ratings of music selections and makes recommendations by finding which selections were liked by multiple people. However, the Ringo system does not take advantage of any available descriptions of the music, such as structured descriptions in a data base, or free text, such as that contained in music reviews. An article titled "Evolving agents for personalized information filtering", published at the Proc. 9th IEEE Conf. on AI for Applications by Sheth and Maes, described the use of agents for information filtering which use genetic algorithms to learn to categorize USENET news articles. In this system, users must define news categories and the users actively indicate their opinion of the selected articles. Their system uses a list of keywords to represent sets of articles and the user profiles are updated using genetic algorithms.

A number of other research groups have looked at the automatic generation and labeling of clusters of articles for the purpose of browsing through the articles. A group at Xerox Parc published a paper titled "Scatter/gather: a cluster-based approach to browsing large article collections" at the 15 Ann. Int'l SIGIR '92, ACM 318-329 (Cutting et al. 1992). This group developed a method they call "scatter/gather" for performing information retrieval searches. In this method, a collection of articles is "scattered" into a small number of clusters, the user then chooses one or more of these clusters based on short summaries of the cluster. The selected clusters are then "gathered" into a subcollection, and then the process is repeated. Each iteration of this process is expected to produce a small, more focussed collection. The cluster "summaries" are generated by picking those words which appear most frequently in the cluster and the titles of those articles closest to the center of the cluster. However, no profiles of users are collected, so no performance improvement occurs over time.

Apple's Advanced Technology Group has developed an interface based on the concept of a "pile of articles". This interface is described in an article titled "A 'pile' metaphor for supporting casual organization of information in Human factors in computer systems" published in CHI '92 Conf. Proc. 627-634 by Mander, R. G. Salomon and Y Wong. 1992. Another article titled "Content awareness in a file system interface: implementing the 'pile' metaphor for organizing information" was published in 16 Ann. Int'l SIGIR '93, ACM 260-269 by Rose E. D. et al. The Apple interface uses word frequencies to automatically file articles by picking the pile most similar to the article being filed. This system functions to cluster articles into subpiles, determine key words for indexing by picking the words with the largest TF/IDF (where TF is term (word) frequency and IDF is the inverse article frequency) and label piles by using the determined key words.

Numerous patents address information retrieval methods, but none develop user profiles based on passive monitoring of which articles the user accesses. None of the systems described in these patents present computer architectures to allow fast retrieval of articles distributed across many computers. None of the systems described in these patents address issues of using such article retrieval and matching methods for purposes of commerce or of matching users with common interests. U.S. Patent No. 5,321,833 issued to Chang et al. teaches a method in which users choose terms to use in an information retrieval query, and specify the relative weightings of the different terms. The Chang system then calculates multiple levels of weighting criteria. U.S. Patent No. 5,301,109 issued to Landauer et al. teaches a method for retrieving articles in a multiplicity of languages by constructing "latent vectors" (SVD or PCA vectors) which represent correlations between the different words. U.S. Patent No. 5,331,554 issued to Graham et al. discloses a method for retrieving segments of a manual by comparing a query with nodes in a decision tree. U.S. Patent No. 5,331,556 addresses techniques for deriving morphological part-of-speech information and thus to make use of the similarities of different forms of the same word (e.g. "article" and "articles").

Therefore, there presently is no information retrieval and delivery system operable in an electronic media environment that enables a user to access information of relevance and

interest to the user without requiring the user to expend an excessive amount of time and energy.

## SOLUTION

The above-described problems are solved and a technical advance achieved in the field by the system for customized information retrieval and delivery in an electronic media environment, which system automatically constructs both an "article profile" for each article in the electronic media based on the frequency with which each word appears in the article relative to its overall frequency of use in all articles, as well as a "user profile" for each user based on which news articles a user is most likely to wish to read. The system then processes the two profiles to generate a user customized rank ordered listing of articles most likely of interest to the user so that the user can select from these relevant articles automatically selected from the plethora of articles available on the electronic media. Because people have multiple interests, multiple profiles are maintained for each user, corresponding to multiple topics of interest. Each user is presented with those articles whose profiles most closely match the user's profiles. User profiles are automatically updated on a continuing basis to reflect each user's changing interests. In addition, articles can be grouped into clusters and menus automatically generated for each cluster of articles to allow users to navigate throughout the clusters and manually locate articles of interest.

In the preferred embodiment of the invention, the system for customized delivery of information uses a fundamental methodology for accurately and efficiently matching users and objects by calculating, using and automatically updating profiles that describe both the users' interests and the objects' characteristics. The objects may be published articles, purchasable items, or even other people. Examples include a person looking for a newspaper story of potential interest, a movie to watch, an item to buy, or another person to correspond with. In all cases, the method is based on determining the similarity between a profile for the object and a profile for the user. Object profiles may include some or all of the following: (1) structured text (as in a newspaper story, a movie review, a product description or an advertisement), (2) structured data (the author of the newspaper article, the actor and directors and rating given to a movie, the price and manufacturer or a product, or the name and phone number of the person placing an advertisement) and (3) a list of the persons who have accessed (read or purchased) the object. The structured data may, in particular, contain

information about the quality of the object, such as measures of its popularity (how often it is accessed) or of user satisfaction (number of complaints received).

The ability to measure the similarity of profiles describing objects and a user's interests can be applied in two basic ways: filtering and browsing. Filtering is useful when large numbers of objects exist in the electronic media space. These objects can be articles which are being received by a user, who only has time to read a small fraction of them. For example, one might potentially receive all items on the AP news wire service, all items posted to a number of news groups, or all advertisements in a set of newspapers, but few people have the time or inclination to read so many articles. A filtering system in the customized information delivery system selects a subset of the articles which the user is more likely to wish to read. The accuracy of this filtering system improves over time by noting which articles the user reads and by generating a measurement of the depth to which the user reads the article. This information is then used to update the generated user's profile.

Browsing provides an alternate method of selecting a small subset of a large number of objects, such as articles. Articles are organized so that users can navigate among the articles by moving from broad groups of articles to smaller, more specific, groups, to individual articles, or from articles or article groups to closely related articles or groups. The methods used by the customized information delivery system allow articles to be grouped into clusters and the clusters to be grouped into larger and larger clusters. These hierarchies of clusters then form the basis for menuing and navigational systems to allow the rapid of search large numbers articles.

There are a number variations on the theme of developing and using profiles for article retrieval, with the basic implementation of an on-line news clipping service representing the preferred embodiment of the invention. Variation of this basic system are disclosed and comprise a system to filter electronic mail, an extension for retrieval of objects such as purchasable items which may have more complex descriptions, a system to automatically build and alter menuing systems for browsing, and a system to construct virtual communities of people with common interests.

## BRIEF DESCRIPTION OF THE DRAWING

Figure 1 illustrates in block diagram form a typical architecture of an electronic media system in which the customized information delivery system of the present invention can be implemented as part of a user server system;

Figure 2 illustrates in flow diagram form the operational steps taken by the customized information delivery system to screen articles for a user;

Figure 3 illustrates in block diagram form additional details of the information server module of the customized information delivery system;

Figure 4 illustrates in flow diagram form a method for automatically generating article profiles and an associated hierarchical menu system;

Figures 5-9 illustrate examples of menu generating process;

Figure 10 illustrates .....

## DETAILED DESCRIPTION

In the preferred embodiment of the invention, the system for customized delivery of information uses a fundamental methodology for accurately and efficiently matching users and objects by calculating, using and automatically updating profiles that describe both the users' interests and the objects' characteristics. The objects may be published articles, purchasable items, or even other people. Examples include a person looking for a newspaper story of potential interest, a movie to watch, an item to buy, or another person to correspond with. In all cases, the method is based on determining the similarity between a profile for the object and a profile for the user. Object profiles may include some or all of the following: (1) structured text (as in a newspaper story, a movie review, a product description or an advertisement), (2) structured data (the author of the newspaper article, the actor and directors and rating given to a movie, the price and manufacturer or a product, or the name and phone number of the person placing an advertisement) and (3) a list of the persons who have accessed (read or purchased) the object. The structured data may, in particular, contain information about the quality of the object, such as measures of its popularity (how often it is accessed) or of user satisfaction (number of complaints received).

The preferred embodiment of the system for customized information retrieval and delivery operates in an electronic media environment for accessing articles, which may be news, electronic mail, other published documents, or product descriptions. The system in its broadest construction comprises three modules, which may be separate entities, or combined into a lesser subset of physical entities. The specific embodiment of this system illustrates the use of a first module which automatically constructs an "article profile" for each article in the electronic media based on the frequency with which each word appears in the article relative to its overall frequency of use in all articles. A second module constructs a "user profile" for each user, which user profile is based on the news articles a user is most likely to wish to read. The system further includes a profile processing module which processes the two profiles to generate a user customized rank ordered listing of articles most likely of interest to the user so that the user can select from these relevant articles automatically selected from the plethora of articles available on the electronic media. Because people have multiple interests, multiple profiles can be maintained for each user, corresponding to multiple topics of interest. Each user is presented with those articles whose profiles most closely match the user's profiles. User profiles are automatically updated on a continuing basis to reflect each user's changing interests. In addition, articles can be grouped into clusters and menus automatically generated for each cluster of articles to allow users to navigate throughout the clusters and manually locate articles of interest.

**Electronic Media System Architecture**

Figure 1 illustrates in block diagram form the overall architecture of an electronic media system in which the system of the present invention can be used to provide user customized access to objects that are available via the electronic media system. In particular, the electronic media system comprises a data communication facility that interconnects a plurality of users with a number of information servers. The users are typically individuals, whose personal computers are connected via modem to a telecommunication network. User information access software is resident on the user's personal computer and serves to communicate via the modem and a telephone connection established in well known fashion over the telecommunication network with a network vendor who provides data interconnection service with selected ones of the information servers. The user can, by use of the user

information access software, interact with the information servers to request and obtain access to data that resides on mass storage systems that are part of the information server apparatus. New data is input to this system by users via their personal computers and by commercial information services by populating their mass storage systems with commercial data. Each user and the information servers have electronic mail addresses which enable data communication connections to be established between a particular user and the selected information server. A user's e-mail address uniquely identifies the user, the user's information access software in an industry standard format such as: username@aol.com or username@netcom.com. The network vendors provide telephone access numbers for their subscribers, through which the users can access the information servers. The subscribers pay the network vendors for the access services on a fee schedule that typically includes a monthly subscription fee and usage based charges. A difficulty with this system is that there are a large number of information servers located around the world, each of whom provide access to a set of information of differing format, content and topics and via a cataloging system that is typically unique to the particular information server. The information is comprised of individual "files" (herein termed "articles"), which can be audio data, video data, graphics data, text data and combinations thereof. The text data can be of the class: commercially provided news articles, published documents, letters, user generated documents, database data or combinations of these classes of data. The organization of the articles containing the information and the native format of the data contained in articles of identical class may vary by information server. Thus, a user can have a difficult task to locate articles that contain the desired information, because the information may be contained in articles, whose information server cataloging may not enable the user to locate the article. Furthermore, there is no standard catalog that defines the presence and services provided by all information servers. A user therefore does not have simple access to information but must expend a significant amount of time and energy to excerpt a segment of the information that may be relevant to the user from the plethora of information that is generated and populated on this system. Even if the user commits the necessary resources to this task, existing information retrieval processes lack the accuracy and efficiency to ensure that the user obtains the desired information.

It is obvious that within the constructs of this electronic media system, the three modules of the customized information delivery system can be implemented in a distributed manner, even with various modules being implemented on and/or by different vendors within the electronic media system. For example, the information servers can include the article profile generation module while the network vendor may implement the user profile generation modules and/or the profile processing module. Various other partitions of the modules and their functions are possible and the example provided represents an illustrative example and is not intended to limit the scope of the claimed invention.

**News Clipping Service**

The customized information delivery system of the present invention can be used in the electronic media system of Figure 1 to implement an automatic news clipping service which learns to select (filter) news articles to match a user's interests, based solely on which articles the user chooses to read. The customized information delivery system generates a profile for each article that enters the electronic media system based on the relative frequency of occurrence of the words contained in the article. The customized information delivery system also generates a profile for each user, as a function of a number of user specific characteristics, which makes each user's profile unique to that user. As new articles are received for storage on the mass storage systems of the information servers, the customized information delivery system generates their profiles, which are compared to the user's profiles, and articles which are closest (most similar) to the user's profile are presented to the user for possible reading. The computer program providing the articles to the user monitors how much the user reads (the number of screens of data and the number of minutes spent reading), and adjusts the user's profiles to more closely match what was read. The details of the method used by this system are disclosed in flow diagram form in Figure 2. This method require selecting a specific method of calculating profiles, of measuring similarity of articles and profiles, and of updating profiles based on what the user read, and the examples disclosed herein are examples of the many possible implementations that can be used and should not be construed to limit the scope of the system.

**Retrieve New Articles From Article Source**

Articles are available on-line from a wide variety of sources. In the preferred embodiment, one would use the current days news as supplied by a news source, such as the AP or Reuters news wire. At step 201 on Figure 2, these news articles are input to the electronic media system by being loaded into the mass storage system of an information server. The article profile module of the customized information delivery system can reside on the information server and, as each article is received by the information server, the article profile module at step 202 generates a profile for the article and stores the profile in an article indexing memory for later use in selectively delivering articles to users. This method is equally useful for selecting which articles to read from electronic news groups and electronic bulletin boards, and can be used as part of a system for screening and organizing electronic mail ('email').

**Calculate Article Profiles**

An article profile is computed for each article to indicate the relative frequency of word occurrences in this article relative to other comparable articles. Many measures can be used, but the preferred profile is calculated using the TF/IDF measure: the term (word) frequency (TF) times the inverse article frequency (IDF), where the word frequency is the number of occurrences of a word in a article divided by the total number of words in the article and the inverse article frequency is taken to be one over the logarithm of the fraction of articles in which the word occurs. Words which occur frequently ("a", "and", "the" ...) can influence calculated article similarities without indicating article topics, and so are typically removed from feature vectors. Other methods of calculating relative word frequencies can also be used, such as latent semantic indexing or probabilistic models. For many applications, it is useful to augment the set of words actually found in the article with a set of synonyms or other words which tend to co-occur with the each word in the article. This provides profiles which match a broader class of articles and reduces the chance of missing desired articles.

Synonym dictionaries which group together similar words can be used to improve the performance of the profiling and menuing systems. Words which are similar can be treated as if they were the same word, or articles containing a word can be augmented with

synonyms of the word or with other words that tend to occur with the word. Thus, an article talking about "staples" could also be matched to articles about "staplers". The stapler example also illustrates the potential utility of morphological analysis where "staple", "stapled" and "staples" would be treated as the same word.

**Compare Article Profiles To User Profiles**

A set of profiles is stored for each user. These can be initially selected by any of a number of procedures. Among the preferred methods are: using the profiles of articles that the user indicates are representative of his interest, asking the user for key words, or using a standard set of profiles for the people in a given demographic mix. The user profile generation module can reside in any of a number of locations in the electronic media system. For example, the user profile can reside on the user's own personal computer and be periodically accessed, such as off hours, by the network vendor as part of a news delivery service. The user profile can reside on the network vendor's apparatus and again be used as part of a news delivery service. Another possibility is that the user's profile is migrated out to a number of selected information servers, which perform all the processing of the customized information delivery service and simply forward desired articles to the user. For the purpose of example, presume that the user profile generation module resides on the network vendor system and is user therein to identify articles, newly stored on various information server systems, that are of interest to the user, whose user profile is presently being processed.

At step 203 of the process, all article profiles of articles newly entered into information server systems, broadcast on the network, or otherwise made available to the network vendor are compared against the selected user's profile. Those articles whose profile most closely correlates to the selected user's profile are selected. The cosign measure of similarity between article profiles and user profile queries can be used for this correlation processing step, and follows the following algorithm:

$$\text{cosine}(\text{Profile}_i, \text{Profile}_j) = \frac{\sum (\text{term}_{ik}\ \text{term}_{jk})}{\quad}.$$

$$\text{sqrt}(\sum \text{term}_{ik}2 \sum \text{term}_{jk}2)$$

where all sums are over k - sums over all the terms (TF/IDFs for each word), $\text{term}_{ik}$ is the TF/IDF for the kth word in profile i, and $\text{term}_{jk}$ is the TF/IDF for the kth word in profile j. Since most words do not appear in most articles, most terms in most article profiles are zero, the articles are substantially orthogonal to each other, and obvious measures of the similarity of between articles such as Euclidean distance do not work well. Other similarity metrics such as the dice measure can also be used.

**Present List Of Articles To User**

Once the user profile/article profile correlation step is completed for a selected user, at step 204 the profile processing module presents a list of titles of the selected articles to the user, who can then select any article for viewing. (If no titles are available, then the first sentences of each article can be used.) The list of article titles is ordered according to the degree of similarity of the article profile to the user's profile. This list is either transmitted in real time to the user as the user is present at their personal computer, or can be transmitted to a user's mailbox, resident on the user's personal computer or within the network server. The user can then elect which if any, of the identified articles the user wishes to review. The user can still access all articles in any information sever to which the user has authorized access, however, those lower on the list are simply further from the users interests, as defined by the generated user's profile.

**Monitor Which Articles Are Read**

The customized information retrieval system at step 205 monitors which articles the user reads, keeping track of how many pages of text are transmitted to the user (read), how much time is spent viewing the article, and whether all pages of the article were viewed. This information can be combined to measure the depth of the user's interest in the article. Although the exact details depend on the length and nature of the articles being searched, a typical formula might be:

measure of article attractiveness =

0.2 if the second page is accessed +

0.2 if all pages are accessed +

0.2 if more than 30 seconds was spend on the article +

0.2 if more than one minute was spend on the article +

0.2 if the minutes spent in the article are greater than half the number of pages

The computed measure of article attractiveness can then be used as a weighting function to adjust the user's profile to thereby more accurately reflect the user's dynamically changing interests.

**Update User Profiles**

Updating of user profiles can be done at step 206 using the method described in copending U.S. Patent Application Serial No. 08/ ???. The user profile closest the article profile of the article read is shifted slightly towards the article profile. Given a set of J articles available with characteristics $\mathbf{d}_{jk}$ (assumed correct for now), and a set of user preferences, $\mathbf{u}_{ik}$ , viewer i would be predicted to pick a set of P articles to minimize:

$$\sum_{\substack{j \text{ in the} \\ \text{best P of J}}} d(\mathbf{d}_{jk}, \mathbf{u}_{ik})$$

The article characteristics $\mathbf{d}_{jk}$ would be some form of word frequencies such as TF/IDF , the user preferences, $\mathbf{u}_{ik}$ , are user profiles, and $d(\mathbf{d}_{jk}, \mathbf{u}_{ik})$ is the distance between them using the cosine measure. If the user picks a different set of P articles than was predicted, the user profile generation module should try to adjust **d** and **u** to more accurately predict the articles the user selected. In particular, **d** and **u** should be shifted to reduce the match on articles that were predicted to be selected but were not selected and also to increase the match on articles

that were predicted not to be selected but were selected. A preferred method is to shift **u** for each wrong prediction for user i and article j using the formula:

$$\mathbf{u}_{ik} = \mathbf{u}_{ik} - e(\mathbf{u}_{ik} - \mathbf{d}_{jk})$$

This adjustment increases the match between a generated user's profile and the article profile of desired articles by making **u** closer to **d** if e is positive - for the case where the algorithm failed to predict aa article that the viewer read. The size of e determines how many example articles one must see to replace what was originally believed. If e is too large, the algorithm becomes unstable, but for sufficiently small e, it drives **u** to its correct value. One could in theory also make use of the fact that the above algorithm decreases the match if e is negative - for the case where the algorithm predicted an article that the user did not read. However, there is no guarantee that **u** will move in the correct direction in that case. If, as is typically the case, there are multiple user profiles which may be applicable, then only the most applicable profile - that with the highest agreement with the article selected - is updated.

One can also shift the term weights **w** using a similar algorithm:

$$\mathbf{w}_{ik} = (\mathbf{w}_{ik} - e|\mathbf{u}_{ik} - \mathbf{d}_{jk}|) \,/\, S_k(\mathbf{w}_{ik} - e|\mathbf{u}_{ik} - \mathbf{d}_{jk}|)$$

This is particularly important if one is combining word frequencies with other, structured data. As before, this increases the match if e is positive - for the case where the algorithm failed to predict an article that the user read, this time by decreasing the weights on those characteristics for which the user profile differs from the profile of the article. Again, the size of e determines how many example articles one must see to replace what was originally believed. Unlike the case for **u**, one also make use of the fact that the above algorithm decreases the match if e is negative - for the case where the algorithm predicted an article that the user did not read. The denominator of the expression assures that the modified weights **w** still sum to one. Both **u** and **w** can be adjusted for each article accessed. When e is small, as it should be, there is no conflict between the two parts of the algorithm.

This description speaks entirely of counting words in an article, but the methods proposed can be trivially extended to counting n-grams of letters. (N-grams are just sequences of N contiguous letters in a article. For example, this sentence contains a sequence of 5-grams which starts "For e", "or ex", "r exa", "exam", "examp", etc. Articles can be clustered based on the similarity of the occurrence of sequences of five letters (or other letter n-grams) in the articles. Clusters are displayed on a two dimensional plot with distances between the articles indicating how dissimilar they are.

In any case, the selected user's profile is updated at step 206 and the process returns to step 203 for the next user in the catalog of users served by the network vendor, in this specific example. Alternatively, when a new article enters the electronic media system, its profile is generated in a timely manner and this profile is inserted into the pool of profiles to identify users who have expressed an interest in this article. There is obviously the need for multiprocessing capability in this system and the exact temporal ordering of the various operations can be determined based on the computing and communications facilities available. Furthermore, the users can pay higher service rates to receive "instant news" so that articles that enter the system are delivered on a rush basis to these users, while other users have their information retrieval requests processed in due course.

**Matching Users For Purchasing And Virtual Communities**

The user profiles described above can also form the basis for matching buyers and sellers, people bartering goods, and people with common interests. These matches may be either pairs of people or corporate entities, as in a buyer and a seller, or they may be larger groups, as in virtual communities of people wishing to discuss subjects of common interest.

It is common for merchandisers to provide catalogues with descriptions of the products which they sell. These are now often available on-line on CD-ROMS or the internet. The profile-based news clipping and information retrieval techniques described above can also be used to help users locate items which they may wish to purchase. Profiles of the product descriptions and of product reviews can be made, treating them as standard articles augmented with structured database information. Users repeatedly shopping for given items can have profiles created and modified to reflect the item descriptions which they find most

interesting. For single purchases, menus generated by clustering items can be used. This is particularly important when a user may be shopping from items sold by a large number of vendors, where vendors' descriptions and product categories are variable.

Such profiles can also be used to match any two people based either on profiles generated from descriptions they provide ("want ads" to buy or to sell) or based on profiles of what they have purchased or sold in the past. Such matching is particularly critical in "thin" markets such as most barters, and in certain contract employment situations, where those involved in the barter or contracting are looking for specific items or abilities. In a market where many different people are looking for unusual products or services, the ability to automatically search through millions of product and service descriptions to find matches is particularly valuable. As described above, the filtering (profile updating) is useful when one is repeatedly seeking similar people (an ongoing need for database consultants with specific experience), while the browsing (cluster-based menu) system is more useful for unusual searches (one is looking for a consultant for a specific data base job, based on a description of that job).

If the descriptions of the purchasable items are free text, then the methods described for the news clipping service can be used directly. However, purchasables often have structured descriptions, including attributes such as price, availability of colors or sizes, delivery times and costs, popularity (volume sold), ratings by users or independent organizations, etc. Such items can be included along with relative word frequencies (TF/IDF measures) in the article profiles, as long as care is taken to accurately compute the relative weightings to be given to the different attributes.

**Virtual Community**

Computer users frequently join other users for discussions on computer bulletin boards or news groups. In current practice, each bulletin board has a specified topic and users then look through a long list of topics (typically hundreds) to find topics of interest. The users then must select for themselves which of thousands of messages they find interesting from among those posted to the selected bulletin boards and, if desired, post additional discussion on the topics. The customized information delivery system described above can also be used to

allow users to find other people with similar interests and to generate virtual communities of people with similar interests. Profiles of each person can be formed by the news-clipping method described above: a person's set of profiles is basically the centroids of the clusters of the profiles of the bulletin board articles read by the user. Users can then be grouped together based on the similarity of one or more of their profiles. Such groups of people have been reading and writing about similar topics and in similar styles and so presumably share interests. New bulletin boards can be formed as sufficient numbers of users with different interests accumulate.

The existence of thousands of Internet bulletin boards (also termed news groups) and countless more private bulletin board services (BBS's) demonstrates the very strong interest among members of the electronic community in forums for the discussion of ideas about almost any subject imaginable. Presently, bulletin board creation proceeds in a haphazard form, usually instigated by a single individual who decides that a topic is worthy of discussion. There are protocols on the Internet for voting to determine whether a news group should be created, but there is a large hierarchy of news groups (all beginning with the prefix "alt.") that do not follow this protocol.

The customized information delivery system can function as a browser for bulletin boards where each bulletin board is characterized by one or more profiles, so that potential new bulletin board users can locate bulletin boards by presenting the customized information delivery system with one or more messages typical of what they are looking for. User profiles are generated from the messages using the methods described above, and then the user identifies those bulletin boards most closely corresponding to the messages. Note that bulletin boards, like all objects with profiles, could also be clustered and put into an automatically generated menu.

The above described method is useful if people with the right set of interests have already formed a bulletin board or chat group. However, because people have varied and varying complex interests, it is desirable to automatically create groups of people ( virtual communities with common interests. The Virtual Community System (VCS) described below is a network-based agent that seeks out users of a network with common interests,

dynamically creates bulletin boards or electronic mailing lists for those users, and introduces them to each other electronically via e-mail.

The functions of Virtual Community Service are general functions that could be implemented on any network ranging from an office network in a small company to the World Wide Web or the Internet. The four main steps in the procedure are:

1. Scanning postings to bulletin boards and clustering people who post similar messages.
2. Create new bulletin boards.
3. Announce new bulletin boards.
4. Enroll users in the bulletin boards.

Each of these steps can be carried out as described below.

**Scanning**

Virtual Community Service locates users with interests in common to form a virtual community. Using the text-searching and clustering technology described above, the Virtual Community Service constantly scans all the news groups and electronic mailing lists on a given network. The network can be the Internet, or a set of bulletin boards maintained by America On-Line, Prodigy, or Compuserve, or a smaller set of bulletin boards that might be local to a single organization, for example a large company, a law firm, or a university.

Clustering the messages sent to bulletin boards based on the similarity of their profiles automatically finds threads of discussion that show common interests among the users. Naturally, discussions on a single bulletin board tend to show common interests; however, this method uses all the texts from every available bulletin board and electronic mailing list. Whenever it finds a cluster of sufficient size (for example, 10-20 different messages) in different bulletin boards or mailing lists, it forms a new virtual community.

**Bulletin Board Creation**

Virtual Community Service creates a bulletin board or an electronic mailing list, whichever is appropriate, representing the newly-formed community. If the newly-found

cluster only contains a small number of members, for example 2-10 people, it is probably most appropriate to create a small mailing list and subscribe these people to it. Virtual Community Service initiates the mailing list by sending an e-mail message containing the text it found and a suggested name (see below for techniques for generating cluster labels) for the new mailing list. After this initial message, users may choose to respond and continue the discussion, or they may let it expire if they prefer. All such mailing lists will be time-stamped, and if they are not used for a user-determined length of time, Virtual Community Service deletes them.

If Virtual Community Service finds a larger number of people engaged in a discussion in different forums, it creates a new bulletin board instead. Virtual Community Service creates a name for this bulletin board and post on it the messages it has collected: all the messages in a cluster. For a short period of time, Virtual Community Service continues collecting messages automatically using its clustering routines, and these messages are posted to the bulletin board. After the time limit has expired, new readers of the bulletin board have to maintain it by posting items voluntarily. Alternatively, the system manager can allow Virtual Community Service to continue posting items automatically for an indefinite period of time, if this is deemed useful.

**Announcing**

Virtual Community Service informs all the members of the new virtual community of the new bulletin board service (or mailing list), and gives each of them a brief summary of the basis of the community. Because bulletin boards are read voluntarily, Virtual Community Service sends an e-mail message directly to everyone who posted any text that served as the basis for a new bulletin board it has created. As stated above, it will do the same for new mailing lists. However, with bulletin boards, it simply informs users of the existence and name of the bulletin board, and leave it to them to decide whether or not they wish to read it.

**Enrolling**

Even after creation of a new Virtual Community, Virtual Community Service continues to scan existing mailing lists and bulletin boards for messages that belong to that community.

Any new messages that appear are cross-posted to the new community, and the person posting the message is informed by Virtual Community Service that they belong to a new Virtual Community. The user can then decide whether or not to read the new Virtual Community Service bulletin board or, if the community is using a mailing list, to subscribe to the mailing list.

With these facilities, Virtual Community Service provides automatic creation of virtual communities in any local or wide-area network. The core technology underlying Virtual Community Service is creating a search and clustering mechanism that can find articles that are "similar" in that the users share interests. This is precisely what was described above. One must be sure that Virtual Community Service does not bombard users with notices about communities in which they have no real interest. On a very small network a human could be "in the loop", scanning proposed virtual communities and perhaps even giving them names. But on larger networks the Virtual Community Service has to run in fully automatic mode, since it is likely to find a large number of virtual communities.

Matching of one person against a second person presents additional technical difficulties, since each person may have multiple profiles. Defining the match between a person with multiple profiles and a article with a single profile is easy: the distance between the person and the article is taken to be the smallest of the distances between the article and each of the person's profiles. For matching pairs or groups of potential virtual community members, one can either 1) match each profile of person A with each profile of person B and take the closest match, in effect determining if A and B have any interests in common or 2) calculate the sum of squares of the distances between the n closest profiles. If n=1, this reduces to (1); otherwise it looks for more areas of overlap. We generally take n to be 2.

**Menu Generation Based On Clustering and Automatic Cluster Labeling**

A browsing system can be constructed for the retrieval of articles, such as reference articles in an on-line library using the same methods underlying the news clipping service. The main difference is that since all reference articles are potentially of interest and not just the most recent ones, as is the case for news, and since users are often looking for specific

topics, the preferred interface includes a menu system that the user can navigate to locate articles of interest to the user.

In order to allow users to rapidly locate a article from among a large set of articles, the customized information delivery system calculates profiles for each article using the TF/IDF word frequency measures described above. Articles are then grouped into clusters using a hierarchical clustering algorithm such as a k-means clustering algorithm. Clusters are groups of articles which are similar to each other, as determined by some uniform metric, such as the cosign measure. Hierarchical clusters produce a tree which divides the articles first into two large clusters of roughly similar articles; then each of these clusters is in turn divided into two or more smaller clusters, which in turn are each divided into yet smaller clusters until a "cluster" is found consisting of a single article. This division of articles provides an efficient method to retrieve a article of interest: the user first chooses between the highest level (largest) clusters and then selects among the smaller clusters. This process is repeated until the user comes to the lowest level in the tree, which are the articles themselves.

Hierarchical trees allow rapid selection of one article from a large set. In ten menu sections from menus of ten items each, one can reach $10^{10}$=10,000,000,000 (one hundred billion) items.

The menu generation system is described in flow diagram form in Figure 4. At step 401, the article profile module calculates article profiles as described above, using a uniform system, such as the TF/IDF measure described above. At step 402, the article profile generation module orders all the articles stored in the mass storage system into a hierarchical cluster. As noted above, articles are clustered into similar groups (with distance defined by the cosine measure) using a hierarchical k-means clustering algorithm. At step 403, the article profile generation module generates labels for each cluster in the hierarchy of clusters that was produced at step 402. Labels are generated for each cluster by finding the center (mean) of the TF/IDF's of the articles in the cluster and selecting those words in the mean of the cluster which are farthest from the average TF/IDF across all articles. The near duplicate words are then removed, with the near duplicate words being words which are morphological variants of the same word (e.g. keep only one of "sale" and "sales" or of "sleep" and "sleeping"). Finally, at step 404, the article profile generation module generates menus from

the hierarchical cluster structure and cluster labels produced at steps 402, 403. The hierarchical clustering gives, as d<*add*>. At step 405, optionally, the system monitors which articles are read and adjusts the generated menus. Each of these steps is described in more detail below.

**Label Generation**

A key requirement to successfully use the menu is the ability to automatically label the clusters. Labels can be generated by selecting the article or articles closest to the center of the cluster and then displaying either the article title, or the set of words in the profile of the centroid of the article cluster which have the highest relative frequency (TF/IDF). More informative labels can be generated by removing redundant synonyms from the labels either using a synonym dictionary or a morphological analyzer (for example: "salt, salted, salty, saltier...." should reduce to "salt") and using terms which have the highest discriminatory ability as labels, rather than those with the highest frequency. *<explain how this is done>* It is also useful to allow users to view the articles as needed in case the above descriptive labels are insufficient.

**Menu Generation**

Although most clustering methods generate binary branching trees, for human users, it is preferable to have 4-8 items in a menu. This is easily accomplished by displaying the four "grandchildren" or eight "grandchildren" of a node in a cluster. (see Figure x.)

**Menu Updating**

As users access articles using the menuing system, certain articles and certain regions of the article profile space are accessed more frequently than others. These regions correspond to the users interests. If, for example, a user frequently accessed articles close to a, b, and d in Figure 5 then the menu in Figure 8 could be modified to show the structure illustrated in Figure 9. A more formal algorithm for this is: associate with each node of the tree a probability that the article the user selects is located under that node, is in that cluster. These probabilities can be either all set equal to the number of articles in the cluster divided by the total number of articles, or, if data are available, can be based on total access to each article by all users. The probability of access of a cluster is the total number of accesses of the article in the cluster divided by the total number of accesses of all articles. Once weightings are known for each cluster at each level of the tree, user menus can be generated to maximize the retrieval efficiency. If menus are chosen so that each menu item has

approximately equal probability of accessing the articles under it, then the user has to go through the minimum number of menus.

*<walk through details>*

**Menu Generation From Article Clusters**

Figure 5 illustrates article profiles. Each letter represents a article. Axes $x_1$ and $x_2$ are two of the hundreds of relative word frequencies. Circles represent clusters. Figure 6 illustrates a hierarchical cluster tree for the articles displayed in Figure 5. Figure 7 illustrates cluster labels for the hierarchical cluster tree shown in Figure 7. Article titles, most distinguishing or most frequent words could also be used for cluster labels. Figure 8 illustrates a collapsed menu tree from the hierarchical cluster tree shown in Figure 7.

**Alternate Embodiments of the Menuing System**

Many variations and improvement on the above menuing system are possible. Different clustering methods can be used, such as fuzzy clustering systems which allow a given article to appear in more than one cluster. When retrieving articles which contain structured information as well as free text, the word frequencies can be supplemented with other terms such as the manually assigned category of the article, the price of the object being described in the article, or expert ratings of the object being described in the article, such as the number of stars given to a movie in a movie review. Similarity measures (distance metrics) can be adjusted to reflect different stress laid by different users on different terms or items.

The automatic menuing system described above can be used in parallel with manually generated classifications of articles when such are available. Users can index into clusters either starting at the top of the tree and moving to more specific subclusters or by starting by giving an English language query which is matched against one of the clusters. After an initial cluster is located by finding the cluster center most similar to the query, the user can move to adjacent clusters. It is generally less efficient to start at the largest cluster and repeatedly select smaller subclusters than it is to write a brief description of what one is looking for and

then to move to nearby clusters if the objects initially recommended are not those desired. <explain further?>

**Clustering**

Consider the following hypothesis: each user reads articles from two different clusters. The goal is to reconstitute the "original clusters" from the observed list of articles read. This can be cast as a clustering problem similar to the k-means, but now the criterion being optimized is a little different:

Sum_i Sum_c (I_iC x_i x_C^bar) ^2

where C is the cluster, i is the article, x_C^bar is the mean of cluster C, and I_iC is an indicator matrix which is zero off the diagonal and is between zero and one on the diagonal, indicating how much the article is in which cluster. Note: one could also use a cosine measure.

Sum_C I_iC = I (the identity)

For k-means, I_iC is either I or 0. For the scenario at the top of the message, I_iC is 0 for all by two clusters, and has a mix of 1's and 0's on the diagonal for the two clusters that the movies do fall in. We have discussed two types of clustering:

1) Object-based clustering, where (a) cluster users based on the similarity of the articles they read or (b) cluster articles based on being read by the same users.

2) Attribute-based clustering, where (a) cluster articles based on the similarity of their attributes (word frequencies) or (b) cluster users based on similar attributes (demographics and psychographics).

One could imagine several different combination methods:

3) Hybrid method 1: create a combined function to be minimized which includes the standard costs associated with some combination of 1a, 1b, 2a and 2b. Simultaneously minimize the distance of articles and users from their cluster centers as found both by both object and attribute clustering, by using standard k-means clustering.

4) Hybrid method 2: do 1b, so that articles are labeled by cluster based on which user read them, then use supervised clustering (maximum likelihood discriminant methods) using the word frequencies to do 2a. This tries to use our knowledge of who read what to do a better job of clustering based on word frequencies. One could similarly combine a1 and b2.

**Summary**

A method has been presented for automatically selecting articles of interest to a user. The method generates profiles of the users based on the relative frequency of occurrence of words in the articles which they read. It is characterized by passive monitoring (users do not need to explicitly rate the articles), multiple profiles per user (reflecting interest in multiple topics) and use of elements of the profiles which are automatically determined from the data (the TF/IDF measure based on word frequencies and descriptions of purchasable items). A method has also been presented for automatically generating menus to allow users to locate and retrieve information on topics on interest. This method clusters articles based on their similarity, as measured by the relative frequency of word occurrences. Clusters are labeled either with article titles or with key words extracted from the article. The method can be applied to large sets of articles distributed over many machines.

The above methods can be applied to anything for which profiles can be generated, these include news articles, reference or work articles, electronic mail, product or service descriptions, people (based on the articles they read or the descriptions of the products they buy), and electronic bulletin boards (based on the articles posted to them). A particular consequence of being able to group people by their interests is that one can form virtual communities of people of common interest, who can then correspond with one another via electronic mail.

**WE CLAIM:**

1. A method for automatically providing a user with access to selected ones of a plurality of articles that are stored on an electronic storage media, where said users are connected by user terminals via data communication connections to an information server system which includes said electronic storage media, said method comprising the steps of:

automatically generating article profiles for articles stored in said electronic storage media, each of said article profiles being generated from the contents of an associated one of said articles; and

enabling access to said plurality of articles stored on said electronic storage media by users via said article profiles.

2. The method of claim 1 wherein said step of enabling access comprises:

correlating a user profile, generated for an identified user, with said generated article profiles to identify ones of said plurality of articles stored on said electronic storage media that are likely to be of interest to said identified user.

3. The method of claim 2 wherein said step of enabling access further comprises:

transmitting a list, that identifies at least one of said identified ones of said plurality of articles, to said identified user; and

providing access to a selected one of said plurality of articles stored on said electronic storage media in response to said identified user selecting an item from said list.

4. The method of claim 3 wherein said step of providing access comprises:

transmitting data, in response to said identified user activating a one of said user terminals to identify said selected item on said list, indicative of said identified user's selection of said selected item from said one user terminal to said information server via a one of said data communication connections.

5. The method of claim 4 wherein said step of providing access further comprises:

retrieving, in response to receipt of said data from said one user terminal, an article identified by said selected item from said electronic storage media; and

transmitting said retrieved article to said one user terminal for display thereon to said identified user.

6. The method of claim 1 wherein said step of enabling access comprises:

automatically generating a user profile for an identified user that is indicative of both article profiles of articles retrieved by said identified user as well as the number of pages of said retrieved articles access by said identified user.

7. The method of claim 6 wherein said automatically generated user profile is also indicative of a length of time said identified user accessed said retrieved articles.

8. The method of claim 1 wherein said step of automatically generating article profiles comprises:

automatically generating a hierarchical menu that directs said users to at least a subset of said plurality of articles stored on said electronic media, comprising:

sorting all articles in said subset into a plurality of clusters of articles based on an empirical measure of similarity of content of said articles, and

generating a hierarchical menu that identifies the content in common of articles sorted into each of said plurality of clusters, to enable said identified user to identify ones of said plurality of articles stored on said electronic storage media that are likely to be of interest to said identified user.

9. The method of claim 8 wherein said step of automatically generating a hierarchical menu further comprises:

ascribing a label to each of said plurality of clusters.

10. The method of claim 8 wherein said step of sorting comprises:

dividing said plurality of articles into at least two clusters based upon said empirical measure of similarity of content of said articles,

subdividing each of said at least two clusters into at least two subclusters based upon said empirical measure of similarity of content of said articles, and

repeating said step of subdividing to produce a multi-level hierarchy of identified clusters.

11. The method of claim 10 wherein said step of generating a hierarchical menu comprises:

ascribing a label to each cluster produced by all steps of dividing and subdividing in said step of sorting.

12. The method of claim 11 wherein said step of ascribing comprises:

identifying at least one term in said generated article profiles produced for ones of said plurality of articles sorted into a cluster that is indicative of the information content of said ones of said plurality of articles sorted into said cluster.

13. The method of claim 11 wherein said step of ascribing comprises:

selecting at least one article of said ones of said plurality of articles sorted into said cluster that are closest to the center of the cluster, and

ascribing a label that is indicative of the information content of said ones of said plurality of articles sorted into said cluster, said label comprising elements of at least one of: a title of said selected at least one article, and a set of words contained in the article profile of said selected at least one article cluster which have the highest relative frequency.

Additional Claims:

1) A method for automatically building profiles of people reading articles in order to retrieve articles of greater interest to the readers. (E.g. a custom news clipping service)

2a) A method for screening email based on comparison of user profiles with those of the mail messages.

4) A method for automatically customizing menu trees to allow users to more rapidly access articles in repeatedly accessed areas of interest.

5) A method for locating products or services based on profiles. Profiles are generated for each product or service based on the word frequencies of words in the product description and reviews and on other descriptive data. Then (a) products or services which the user often seeks information about can be selected for the user as in claim 1 or (b) products and services can be clustered and compiled into a menu as in claims 2-4.

6) A method for matching up different people with common interests for purposes such as buying, selling or bartering goods or services. Profiles are generated for each individual and then individuals are matched based on similarity of their profiles. Again, either a) individuals similar to those who the user often seeks information from can be "clipped" for the user as in claim 1 or (b) individuals can be clustered and compiled into a menu as in claim 2-3.

7) A method for matching of sets of people with common interests ("virtual communities") based on profiles developed from the messages which the people read and send.

8) A saleable method for retrieving articles distributed over large numbers of computers. <to be completed by JAMS>

9) A method for combining information on what articles each user has retrieved with demographic or other descriptions of the users and with attributes of the article such as word frequencies in order to more accurately cluster articles into groups. These groups can then be used both for article retrieval and for article filtering.

# SYSTEM FOR CUSTOMIZED INFORMATION DELIVERY

## ABSTRACT

The system for customized information retrieval and delivery in an electronic media environment, which system automatically constructs both an "article profile" for each article in the electronic media based on the frequency with which each word appears in the article relative to its overall frequency of use in all articles, as well as a "user profile" for each user based on which news articles a user is most likely to wish to read. The system then processes the two profiles to generate a user customized rank ordered listing of articles most likely of interest to the user so that the user can select from these relevant articles automatically selected from the plethora of articles available on the electronic media. Because people have multiple interests, multiple profiles are maintained for each user, corresponding to multiple topics of interest. Each user is presented with those articles whose profiles most closely match the user's profiles. User profiles are automatically updated on a continuing basis to reflect each user's changing interests. Alternatively, articles are grouped into clusters and menus are automatically generated for each cluster of articles to allow users to navigate throughout the clusters and manually locate articles of interest.

**This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record**

## BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☒ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☒ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

**IMAGES ARE BEST AVAILABLE COPY.**
**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**